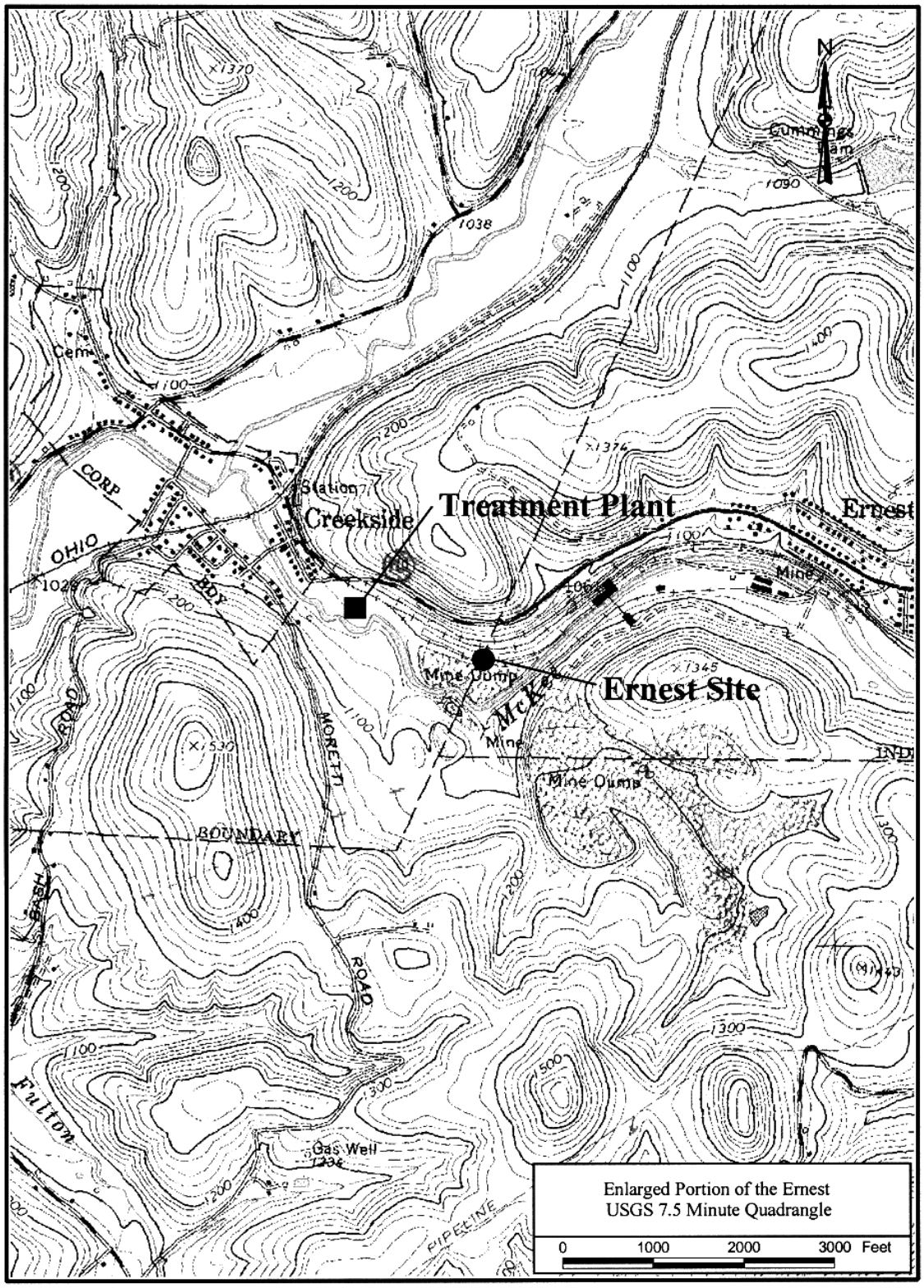
Chapter 6: Analysis of Data from Ernest Refuse Pile Site, Indiana County, PA

The Ernest mine site is located in the Crooked Creek watershed in Indiana County, PA near the town of Ernest (see Figure 6.0). The U.S. Army Corps of Engineers (USACE) completed construction of Crooked Creek dam in 1940 and has managed the lake since then for flood control and recreational purposes. The Commonwealth of Pennsylvania constructed and operated Crooked Creek State Park at the lake prior to 1981 when the USACE acquired the facility. Some portions of the Crooked Creek watershed were impacted by acid mine drainage from extensive bituminous coal mining, particularly the McKee Run tributary, from the town of Ernest downstream to the town of Creekside at the confluence with the main stem of Crooked Creek. The Ernest mine complex, including a large underground mine and associated coal refuse pile, was operated from the early 1900's to 1965 when the mine was abandoned.

An acid mine drainage treatment plant was constructed by the Pennsylvania Department of Environmental Protection and operated from June 1978 until May 1980 when problems with iron sludge recycling operations led to the closure of the plant. The water quality samples and flow measurements from the Ernest refuse pile discharge that are discussed in this chapter were collected between March 1981 and December 1985 as part of studies to evaluate water quality and aquatic biology in the Crooked Creek watershed following closure of the treatment plant. The raw data are listed in Appendix D. There are 198 observations (N = 198), consisting of values for 10 parameters: 1) Days (developed from the date that the sample was taken); 2) pH; 3) Flow; 4) Acidity; 5) Acid load; 6) Total Iron (Fe); 7) Total Iron load; 8) Ferrous Iron (FFe); 9) Sulfate ($S0_4$); 10) Sulfate load.

There is a rather large time gap (four months) between the first three observations and the remainder of the samples that were collected at approximately weekly intervals. There were also at least 15 samples without pH and/or ferrous iron data. After these samples were omitted and other adjustments were made (see Figure 3.1), a revised data set of 174 observations was compiled and used for most of the statistical analyses presented in this chapter. Time gaps in the data should be considered in examining the time series analyses, because elements of the time series analysis assume that there are equal intervals between observations.

Figure 6.0: Map of Ernest Mine Site



Univariate Analysis

Summary statistics for the adjusted data (N = 174) are presented in Table 6.1.

Table 6.1: Summary Statistics of Data (N=174)

	N	Mean	Median	Trimmed Mean	Standard Deviation	Standard Error of the Mean
Days	174	985.6	1027.0	995.3	457.2	34.7
pH	174	2.5061	2.5000	2.5018	0.1524	0.0116
Flow	174	127.2	51.0	85.9	337.0	25.5
Acidity	174	3621	3539	3585	1357	103
Acid Load	174	3367	1843	3031	3639	276
Total iron	174	527.2	515.5	520.5	210.0	15.9
Iron Load	174	626.6	275.0	563.1	722.3	54.8
Ferrous Iron	174	364.8	360.5	351.5	251.9	19.1
SO_4	174	3887.4	3804.0	3915.5	1105.2	83.8
SO_4 Load	174	3837	2108	3431	4198	318

	Minimum	Maximum	First Quartile	Third Quartile	Coefficient of Variation
Days	0.0	1735.0	703.0	1343.5	46.4
pH	2.1	3.1	2.4	2.6	6.1
Flow	2.0	3188.0	8.0	163.2	265.0
Acidity	778	16401	3016	4301	37.5
Acid Load	111	17663	412	5641	108
Total iron	20.0	1929.0	395.0	653.0	39.8
Iron Load	10.0	2758.0	50.5	1147.7	115
Ferrous Iron	8.0	1760.0	161.5	512.0	69.1
SO_4	142.0	6115.0	3155.0	4759.5	28.4
SO_4 Load	117	17746	513	6394	109

The coefficient of variation (CV%) remains within fairly reasonable limits for pH, acidity, and iron. However, variability in ferrous iron (69%) is large. Sulfate is in reasonable control (CV = 28%). Flow, acid load, iron load, and sulfate load show very large variability (all greater than CV = 100%) which suggests that the large variability of the load-type variables is largely due to the high degree of variability shown by flow. These parameters require log transformation to control this variability. The frequency distribution of pH is symmetrical (Figure 6.1a) while flow is skewed, although the major part of the skewness arises from two extremely high values (Rows 142-3 in Appendix Table D, flow = 3003.0 gpm and 3188.0 gpm respectively). All other values of this parameter range from less than 10 to hundreds. Similarly, acidity has an extremely high value (Appendix D Table, Row 143 = 16,401 mg/L); acid load (Figure 6.1b) is skewed.

Total iron and total iron load follow the same pattern. Ferrous iron has one exceptional value. Sulfate (Figure 6.1f) is negatively skewed, whereas sulfate load is positively skewed. This behavior is an indication of the effect that flow can have on a parameter.

Table 6.2: Summary Statistics of Data (N=174)

	N	Mean	Median	Trimmed Mean	Standard Deviation	Standard Error of the Mean
Days	174	985.6	1027.0	995.3	457.2	34.7
pH	174	2.5061	2.5000	2.5018	0.1524	0.0116
Log Flow	174	1.6062	1.7076	1.6081	0.6970	0.0528
Log Acidity	174	3.5349	3.5489	3.5440	0.1466	0.0111
Log Acid Load	174	3.1854	3.2654	3.1930	0.6138	0.0465
Log Total Iron	174	2.6836	2.7122	2.6997	0.2066	0.0157
Log Iron Load	174	2.3564	2.4393	2.3703	0.7244	0.0549
Log Ferrous Iron	174	2.3989	2.5569	2.4442	0.4696	0.0356
Log SO_4	174	3.5646	3.5802	3.5815	0.1748	0.0133
Log SO_4 Load	174	3.2403	3.3240	3.2480	0.6143	0.0466

	Minimum	Maximum	First Quartile	Third Quartile	Coefficient of Variation
Days	0.0	1735.0	703.0	1343.5	46.4
pH	2.1000	3.1000	2.4000	2.6000	6.1
Log Flow	0.3010	3.5035	0.9031	2.2127	43.4
Log Acidity	2.8910	4.2149	3.4795	3.6336	4.1
Log Acid Load	2.0453	4.2471	2.6149	3.7514	19.3
Log Total Iron	1.3010	3.2853	2.5966	2.8149	7.7
Log Iron Load	1.0000	3.4406	1.7032	3.0598	30.7
Log Ferrous Iron	0.9031	3.2455	2.2082	2.7093	19.6
Log SO_4	2.1523	3.7864	3.4990	3.6776	4.9
Log SO_4 Load	2.0682	4.2491	2.7098	3.8058	18.9

Summary statistics for log (base 10) transformed data are listed in Table 6.2 (N = 174). The variables are now either well-behaved (CV $\leq$ 20%) or are not too extreme (CV $\leq$ 50%). Load variables show the largest CV%. This is most likely largely due to flow variability.

Histograms of the log transformed data are displayed in Figures 6.1c, 6.1e, and 6.1g. By plotting the histograms of the original data alongside that of the transformed data, the effect of the transformation is clear. Because pH is already expressed in logarithms, no transformation was applied. In all other parameters, log transformation expanded low magnitude values and reduced asymmetry (for acid load in Figures 6.1b and 6.1c), sometimes perhaps, too much (Figures 6.1d

and 6.1e, iron and log iron respectively). Similarly, because the histogram of sulfate is negatively skewed, log transformation accentuated the negative skewness (Figures 6.1f and 6.1g) making log transformation unnecessary. All load variables are strongly positively skewed when untransformed and the log transformation helps to improve their symmetry.

Figure 6.1a: Histogram of pH, (N = 174)

```
                Histogram of PH   N = 174
                Each * represents 2 obs.
Midpoint    Count
     2.1        1  *
     2.2        3  **
     2.3       28  **************
     2.4       23  ************
     2.5       48  ************************
     2.6       51  *************************
     2.7       14  *******
     2.8        3  **
     2.9        1  *
     3.0        0
     3.1        2  *
```

Figure 6.1b: Histogram of Acid Load, (N = 174)

```
Histogram of C5   N = 174
Each * represents 2 obs.

Midpoint    Count
       0       70  ************************************
    2000       35  ******************
    4000       18  *********
    6000       18  *********
    8000       14  *******
   10000       11  ******
   12000        6  ***
   14000        1  *
   16000        0
   18000        1  *
```

Figure 6.1c: Histogram of Log Acid Load (N=174)

```
Histogram of ACIDLD   N = 174

Midpoint    Count
     2.0        3  ***
     2.2       12  ************
     2.4       19  *******************
     2.6       15  ***************
     2.8       16  ****************
     3.0       12  ************
     3.2       13  *************
     3.4       16  ****************
     3.6       17  *****************
     3.8       24  ************************
     4.0       25  *************************
     4.2        2  **
```

Figure 6.1d: Histogram of Total Iron (N=174)

```
Histogram of FE   N = 174
Each * represents 2 obs.

Midpoint    Count
       0        3  **
     200       11  ******
     400       70  ************************************
     600       60  ******************************
     800       27  **************
    1000        1  *
    1200        1  *
    1400        0
    1600        0
    1800        0
    2000        1  *
```

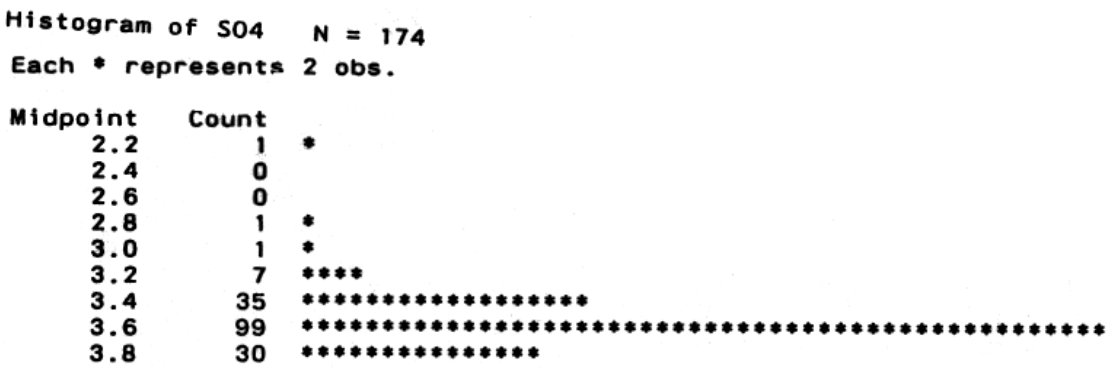
Figure 6.1e: Histogram of Log Total Iron (N=174)

```
Histogram of FE   N = 174
Each * represents 2 obs.

Midpoint    Count
     1.4        1  *
     1.6        0
     1.8        0
     2.0        4  **
     2.2        0
     2.4       15  ********
     2.6       65  ************************************
     2.8       73  ****************************************
     3.0       15  ********
     3.2        1  *
```

Figure 6.1f: Histogram of SO_4 (N=174)

```
Histogram of SO4   N = 174

Midpoint    Count
       0        1  *
     500        0
    1000        2  **
    1500        2  **
    2000        6  ******
    2500       10  **********
    3000       27  ****************************
    3500       38  ****************************************
    4000       19  ********************
    4500       26  ***************************
    5000       22  ***********************
    5500       15  ***************
    6000        6  ******
```

Figure 6.1g: Histogram of Log SO_4 (N=174)



Bivariate Analysis

The bivariate statistical analysis of the Ernest data includes bivariate plots (routinely used in regression and correlation analyses), the use of a correlation matrix to compare and evaluate correlation coefficients, and the use of cross correlation functions to determine if lags in the data for certain parameters tend to obscure correlations that may be present. The correlation matrix is an element of some multivariate statistical analyses, such as principal components analysis and factor analysis (in the r mode). The cross-correlation function is an element of time series analysis because it computes and graphs correlations between two time series. Both of these statistical tools are included in this discussion of bivariate analysis because they are useful in examining the relationship between pairs of variables.

The correlation coefficients for all pairs of variables are shown in Table 6.3. The correlation coefficient (r) at the five percent probability level is given above the table and all correlation coefficients larger than this number are significantly different from zero. For example, only iron vs. pH (r = 0.124) is not significantly different from zero. Similarly, ferrous iron vs. acidity (r = 0.045) and sulfate vs. ferrous iron (r = 0.083) are also not significantly different from zero. All other coefficients reflect a real association (statistically significant), however, in many cases, the degree of association (r^2 x 100%) is small. For example, the correlation of acidity and pH (r = −0.365) indicates an inverse linear association between the two variables as would be expected, but the degree of association is small (r^2 = 13%).

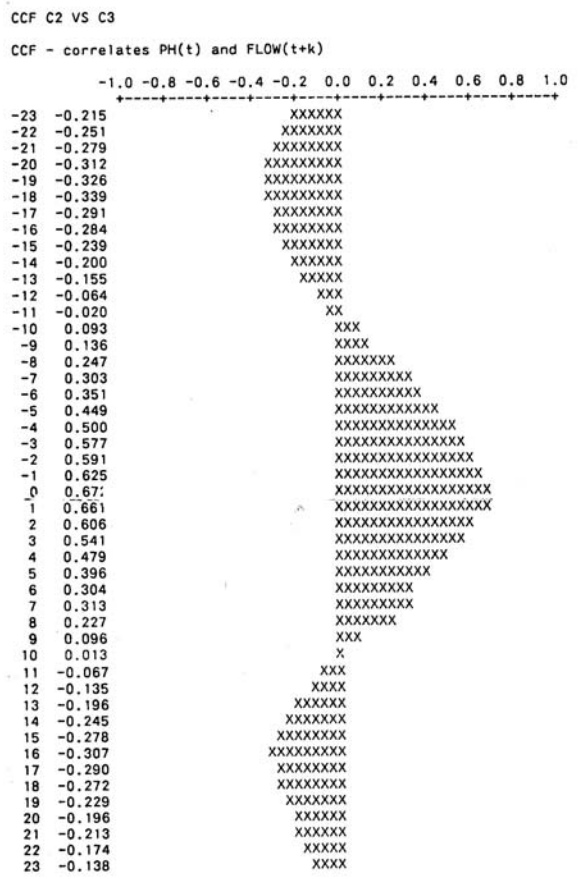
Table 6.3: Correlation Coefficients for 9 Parameters (N=174, $r_{0.05}$ = 0.159)

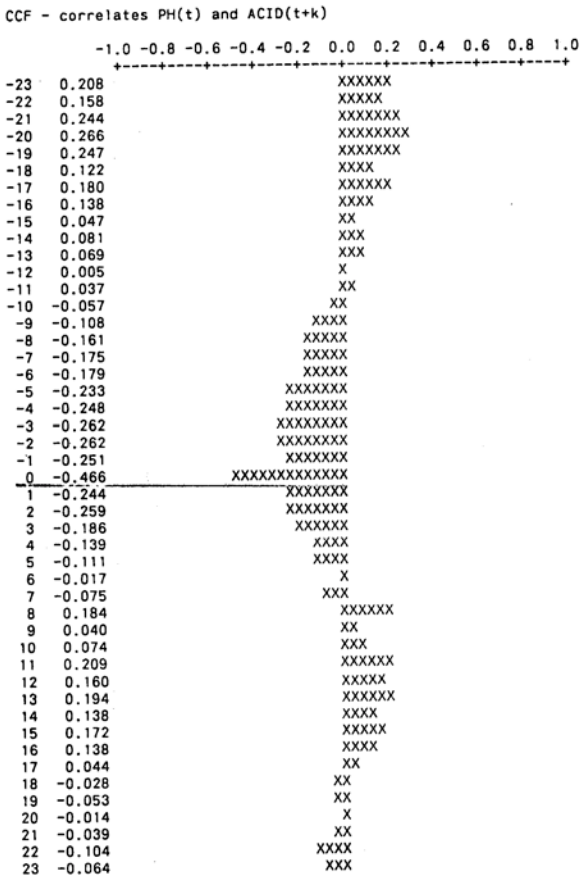
	pH	Flow	Acid	Acid Load	Total Iron	Iron Load	Ferrous Iron	SO_4
Flow	0.191							
Acidity	−0.365	0.308						
Acid Load	0.483	0.206	−0.224					
Total Iron	0.124	0.337	0.526	0.263				
Iron Load	0.498	0.229	−0.262	0.913	0.375			
Ferrous iron	0.248	−0.020	0.045	0.337	0.480	0.388		
SO_4	−0.547	−0.184	0.600	−0.307	0.174	−0.339	0.083	
SO_4 Load	0.472	0.438	−0.030	0.906	0.386	0.890	0.285	−0.293

There are three large correlation coefficients between acid load vs. iron load, acid load vs. sulfate load, and iron load vs. sulfate load. These correlation coefficients are all around r = 0.9 (i.e., about 80 percent in common), probably because of the domination of flow in the measurement of load variables. Whereas, the individual concentration variables acidity vs. iron (r = 0.526), acidity vs. sulfate (r = 0.6), and iron vs. sulfate (r = 0.174) show much lower association (the largest r^2 is 36 %). In addition, any load variable vs. concentration of the same variable shows no appreciable relationship. Thus, the relatively high correlation coefficients due to the inclusion of flow in all load variables is an artifact from the calculation for load (concentration x 0.01212 x flow).

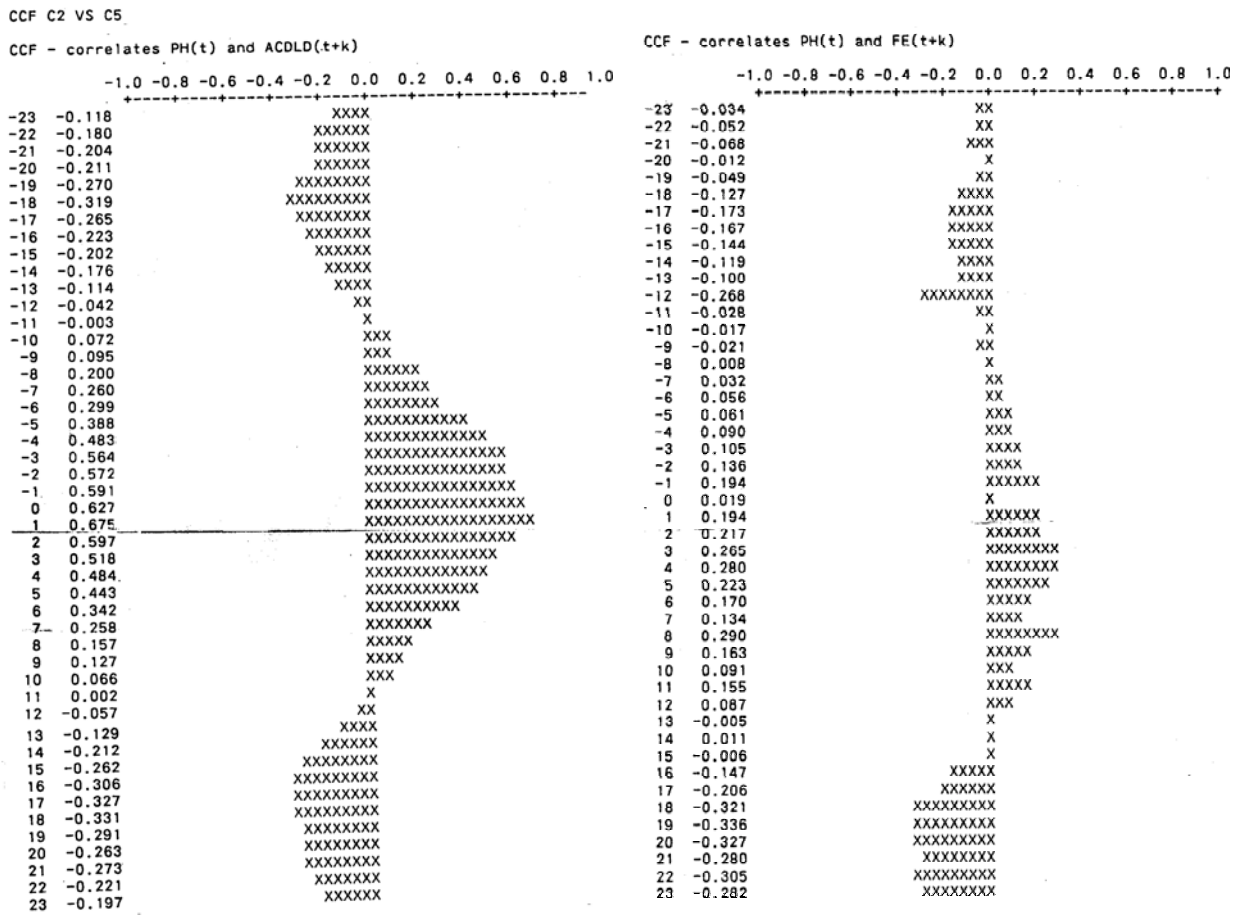
When one examines the cross-correlation functions (Figures 6.2a to 6.2d), it can be seen that the largest correlation occurs at lag zero in Figure 6.2a (pH vs. log flow) and at lag one in Figure 6.2c (pH vs. log acid load) and that the correlations are of the same order of magnitude. Because pH vs. log acidity (Figure 6.2b) yields the strongest r = −0.466 at lag zero, which is much weaker than the value yielded by pH vs. acid load (Figure 6.2c), it is suspected that the effect of flow on load is responsible for the higher correlation. The highest correlation in Figure 6.2d (pH vs. log iron) occurs at lag 19 (r = −0.336), but values of r > 0.25 occur haphazardly at many lags and any association is likely to be very weak.

Figures 6.2a and 6.2b: Cross Correlation Functions of pH vs. Log Flow, and pH vs. Log Acid (respectively)





Figures 6.2c and 6.2d: Cross Correlation Functions of pH vs. Log Acid Load, and pH vs. Log Iron (respectively)



When either pH (which is a logarithmic measure) or logarithms of the other parameters are plotted against days, they appear to show periodic variation with a very large degree of scatter (see for example, pH vs. days (Figure 6.3a) and log flow vs. days (Figure 6.3b)). Log acidity vs. days was not as evident, but log acid load vs. days (Figure 6.3c) is clearly periodic. Here again, the effect of flow on load is likely to be responsible for the cyclical appearance.

Figure 6.3a: Plot of pH vs. Time (days)

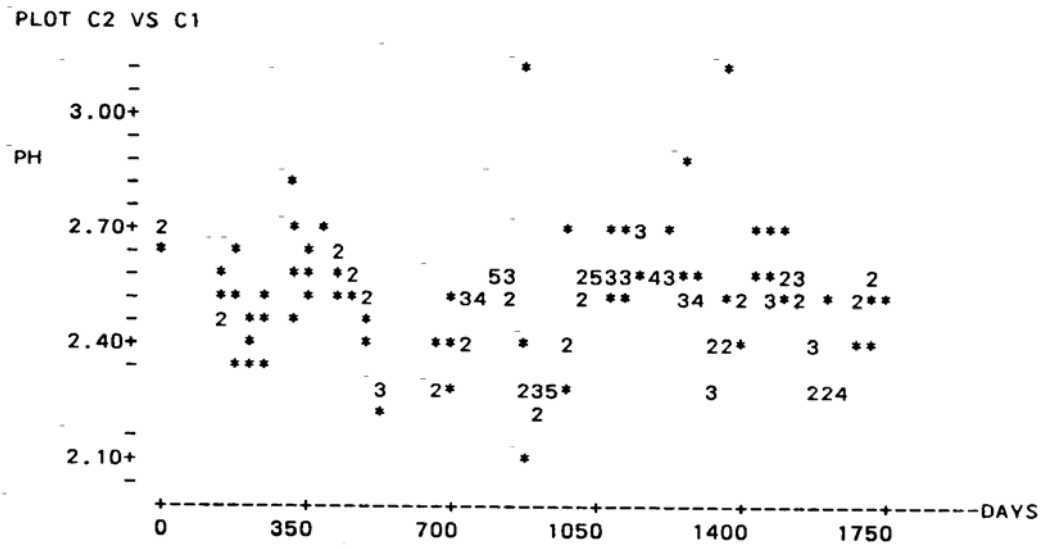


Figure 6.3b: Plot of Log Flow vs. Time (days)

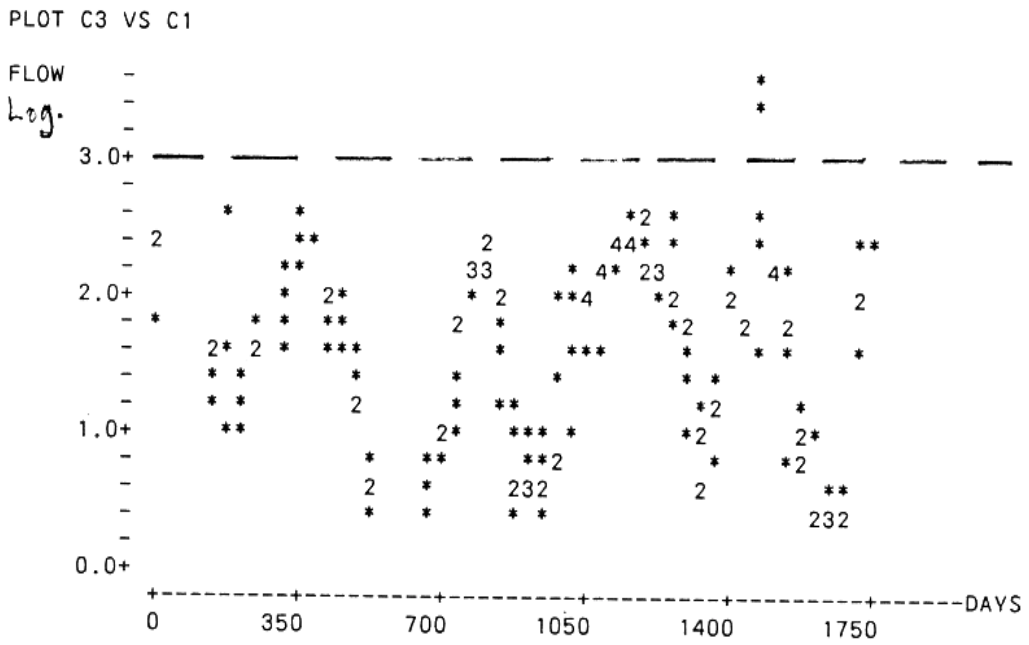
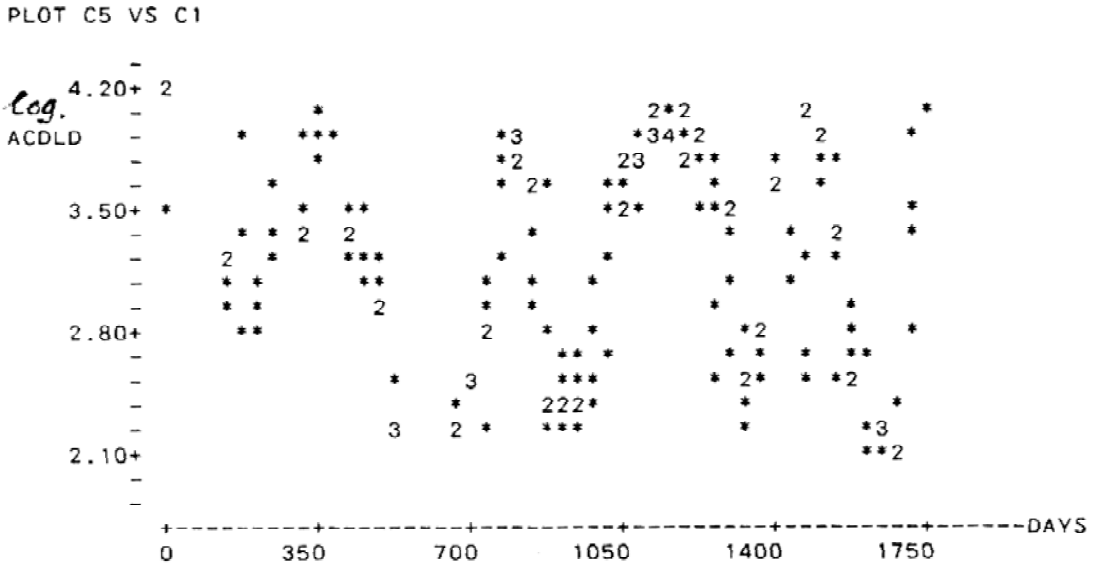
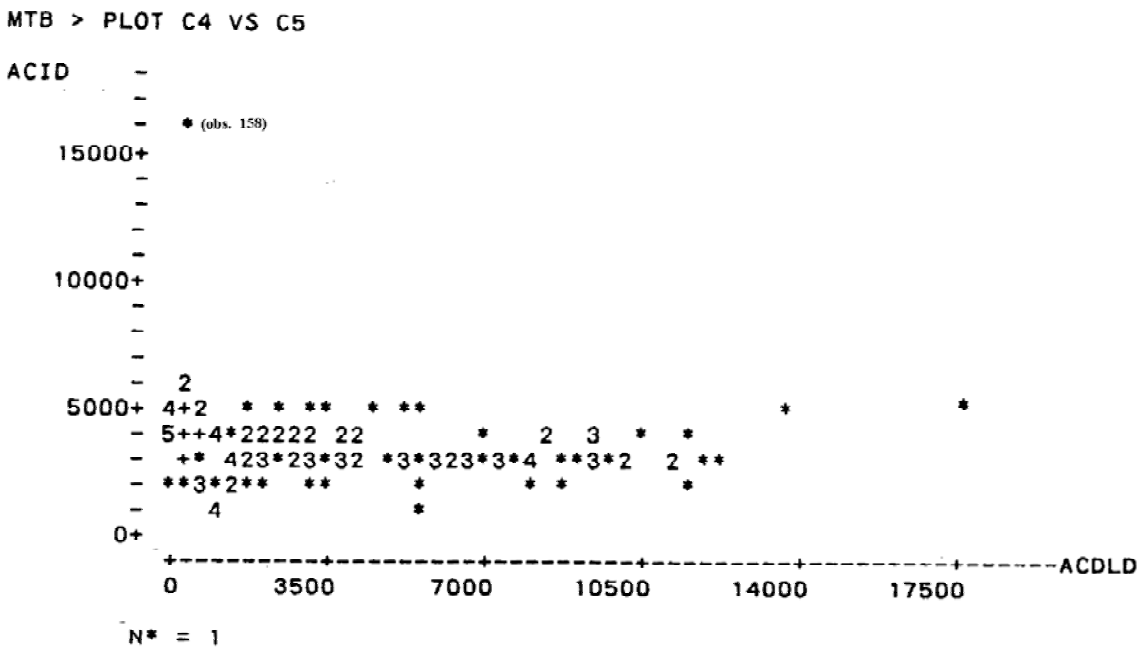


Figure 6.3c: Plot of Log Acid Load vs. Time (days)



Bivariate plots of untransformed data were made and it was found that in most cases, there was little relationship between concentration and load (e.g., Figure 6.4, acidity vs. acid load). The only discrepancies are extreme values which occur as outliers (e.g., observation 158).

Figure 6.4: Plot of Acid vs. Acid Load



bivariate plots of untransformed load variables are included in Figures 6.5a through 6.5c. In Figures 6.5a and 6.5c, the spread of the variables increases with magnitude (i.e., the data are heteroscedastic and so should be expressed in logarithms). Figure 6.5b (acid load and sulfate load) is reasonably homoscedastic, indicating that sulfate load and acid load are not skewed in their frequency distribution. There are obvious extreme outliers in each of the three figures (e.g., observation 133 in Figures 6.5a and 6.5b, and observation 158 in Figures 6.5b and 6.5c).

Figure 6.5a: Plot of Iron Load vs. Acid Loading

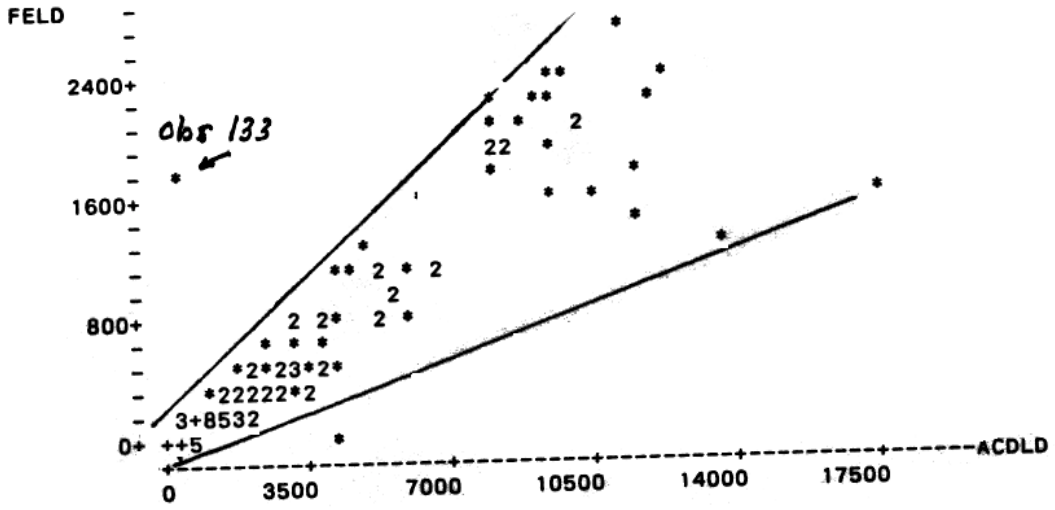


Figure 6.5b: Plot of Acid Load vs. Sulfate Loading

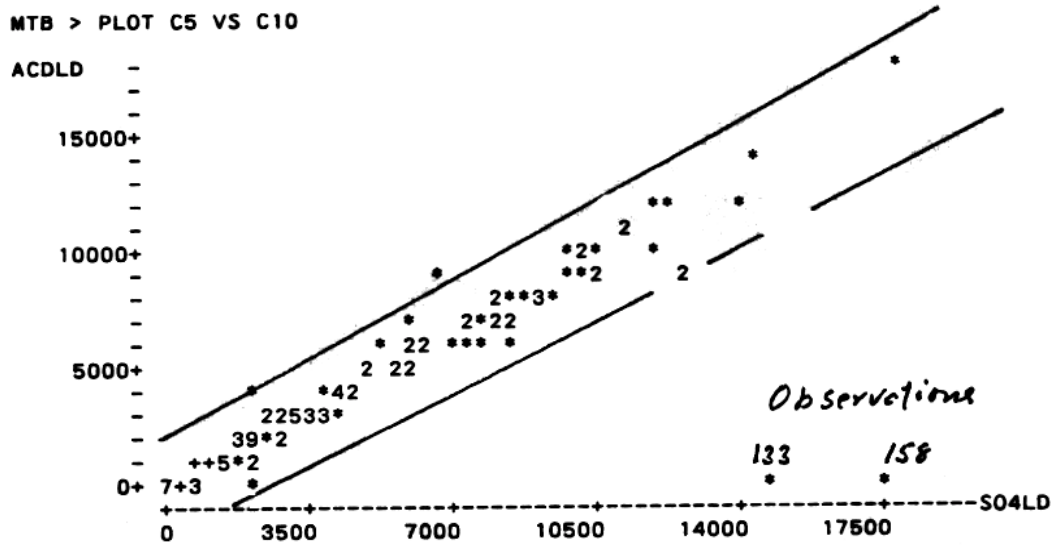
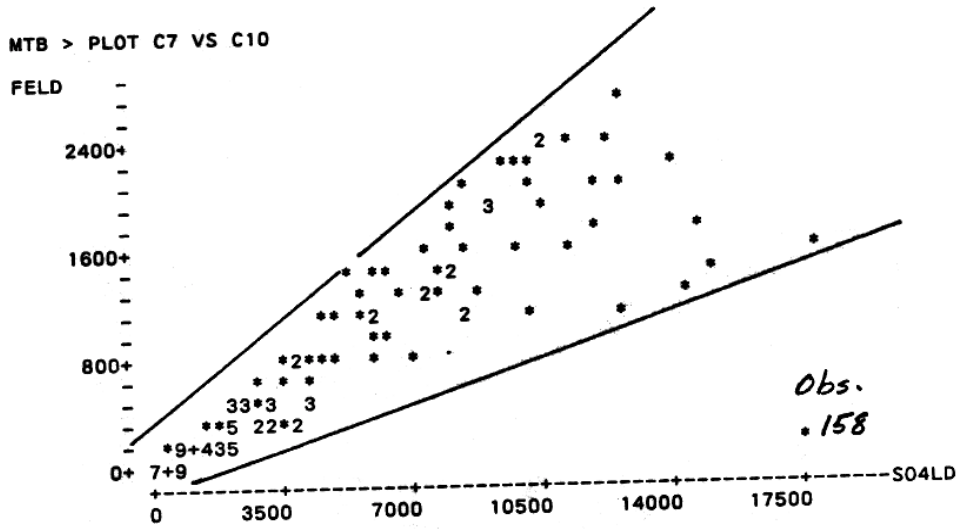


Figure 6.5c: Plot of Iron vs. Sulfate Load



Bivariate plots of logarithmically transformed data are shown in Figures 6.6a to 6.6d. Log acidity vs. log flow (Figure 6.6a) shows no relationship. The exceptional values of two observations of flow occur as outliers. Log acid load, iron load, and sulfate load vs. log flow showed strong linear associations (Figure 6.6b), with various outliers for the extreme values of flow. There appears to be no simple relationship between log acidity and log acid load (Figure 6.6c). The only real association appears to be positive linear between log sulfate and log acid (Figure 6.6d) which, as would be expected, tend to increase together. The presence of two extreme outliers probably would diminish the value of the correlation coefficient between them.

Figure 6.6a: Bivariate Plot of Log Acidity vs. Log Flow

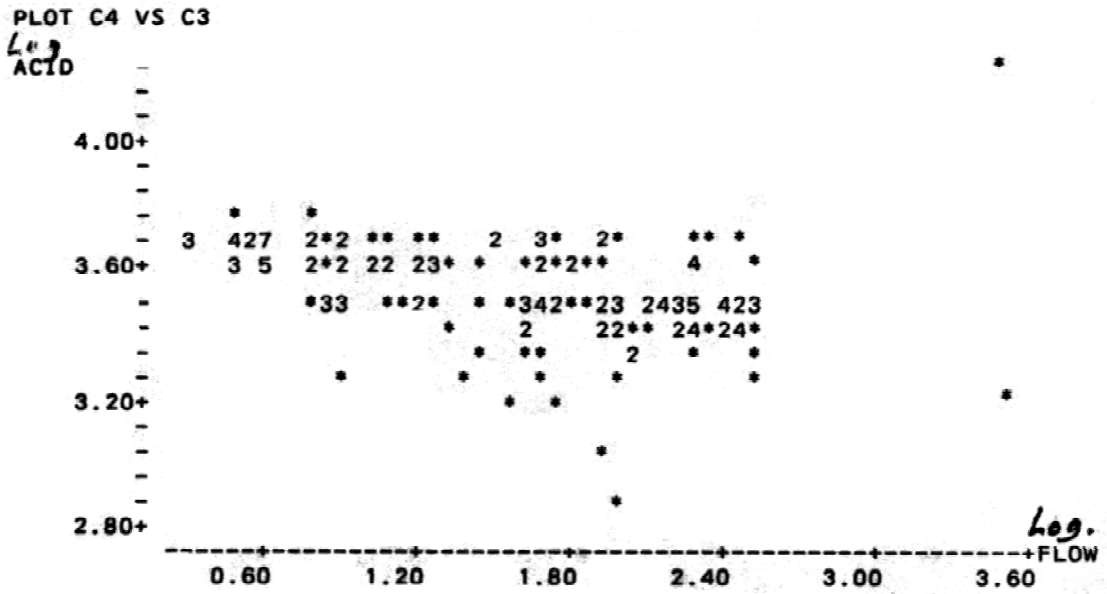


Figure 6.6b: Bivariate Plot of Log Sulfate vs. Log Flow

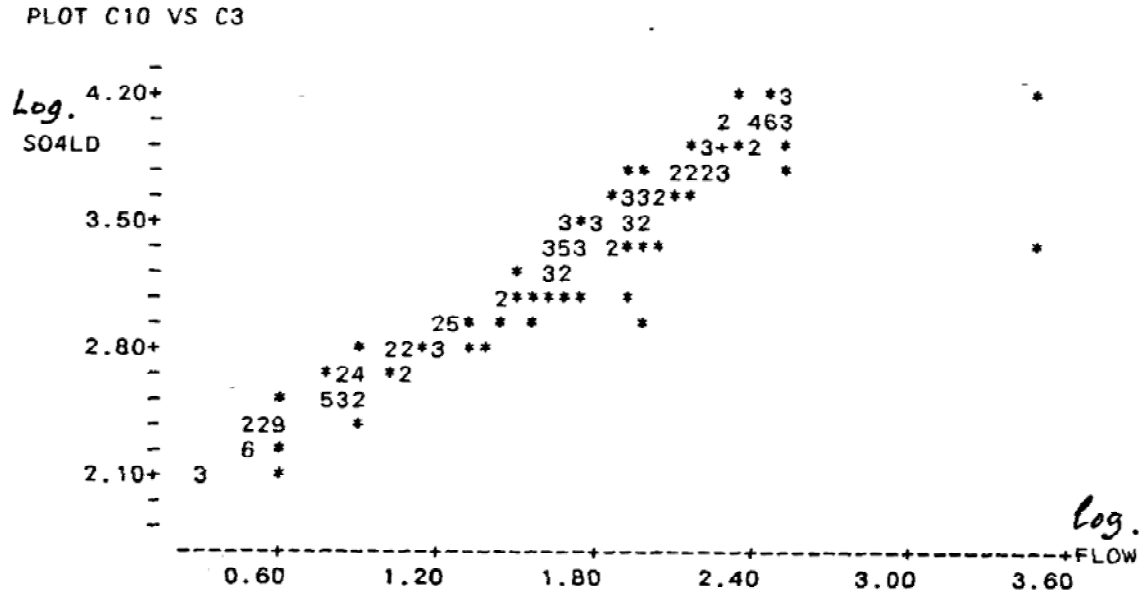


Figure 6.6c: Bivariate Plot of Log Acidity vs. Log Acid Load

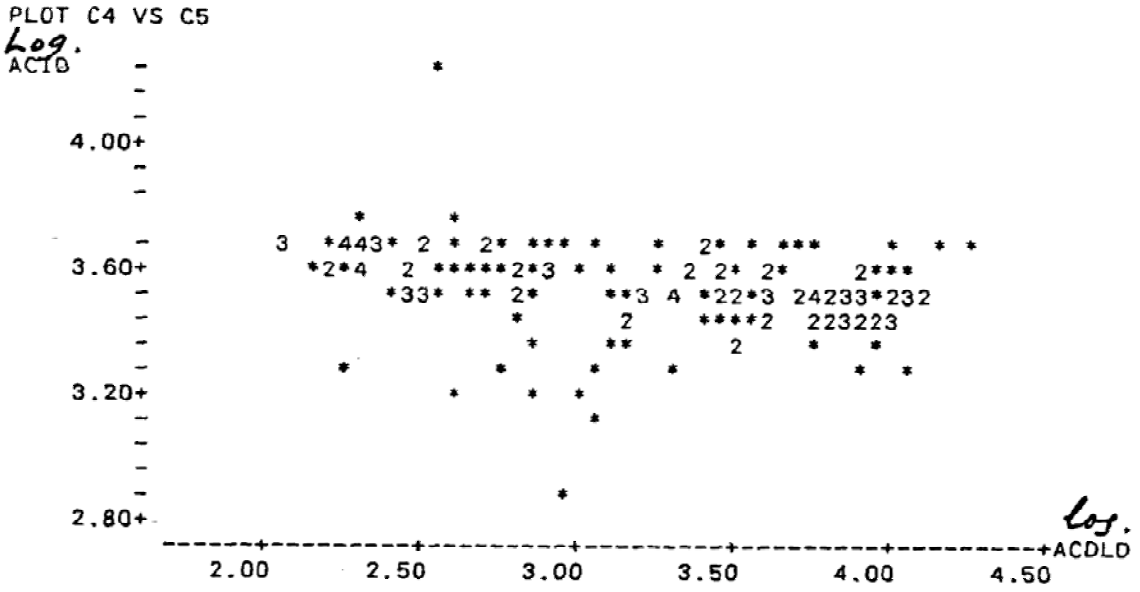
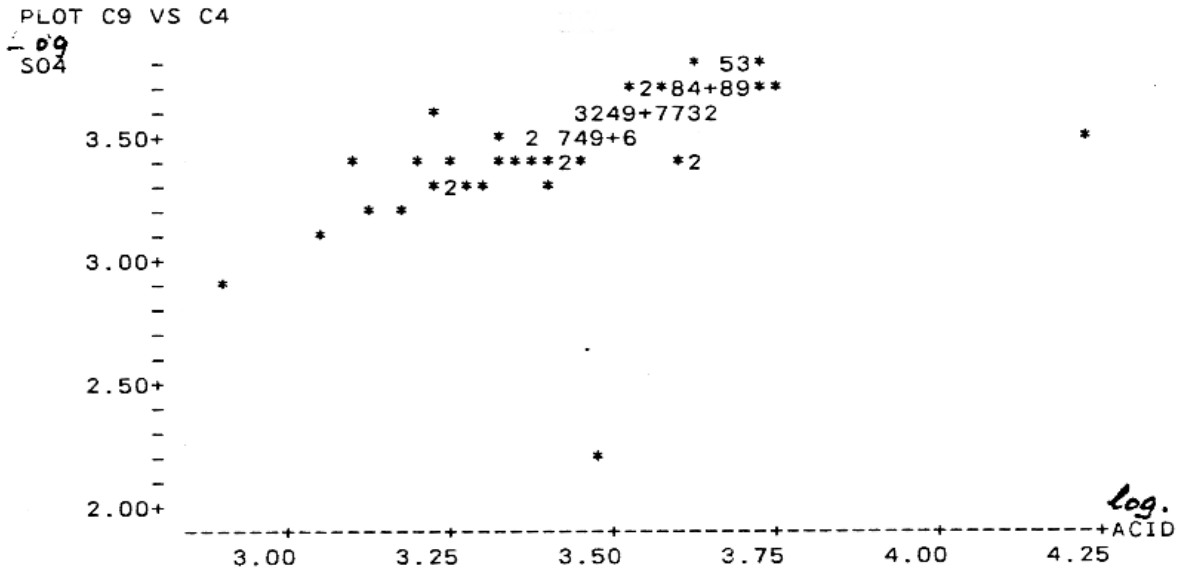


Figure 6.6d: Bivariate Plot of Log Sulfate vs. Log Acid



Time Series Analysis

Time series plots of six selected variables are displayed in Figures 6.7a through 6.7f. pH (Figure 6.7a) illustrates the gap of missing data (September through December, 1982) and possesses two extreme positive values during July 1983 (pH = 3.1) and December 1984 (pH = 3.1). The July 1983 maximum is followed by an extreme minimum (pH = 2.1). Time series plots of flow (Figure 6.7b) and acidity (Figure 6.7c) are dominated by extreme values (March 19 and 26 for the former, and March 26 for the latter).

Time series plots of the load variables (iron, acid and sulfate Figures 6.7d, 6.7e, and 6.7f respectively) are similar and appear to possess a seasonal component in May of each year. This apparent cyclicity is confounded by maxima in March and September 1981, August 1984, and April 1985. The most striking feature is the remarkable similarity in all three graphs, a feature not evident in graphs of the variables expressed as concentrations.

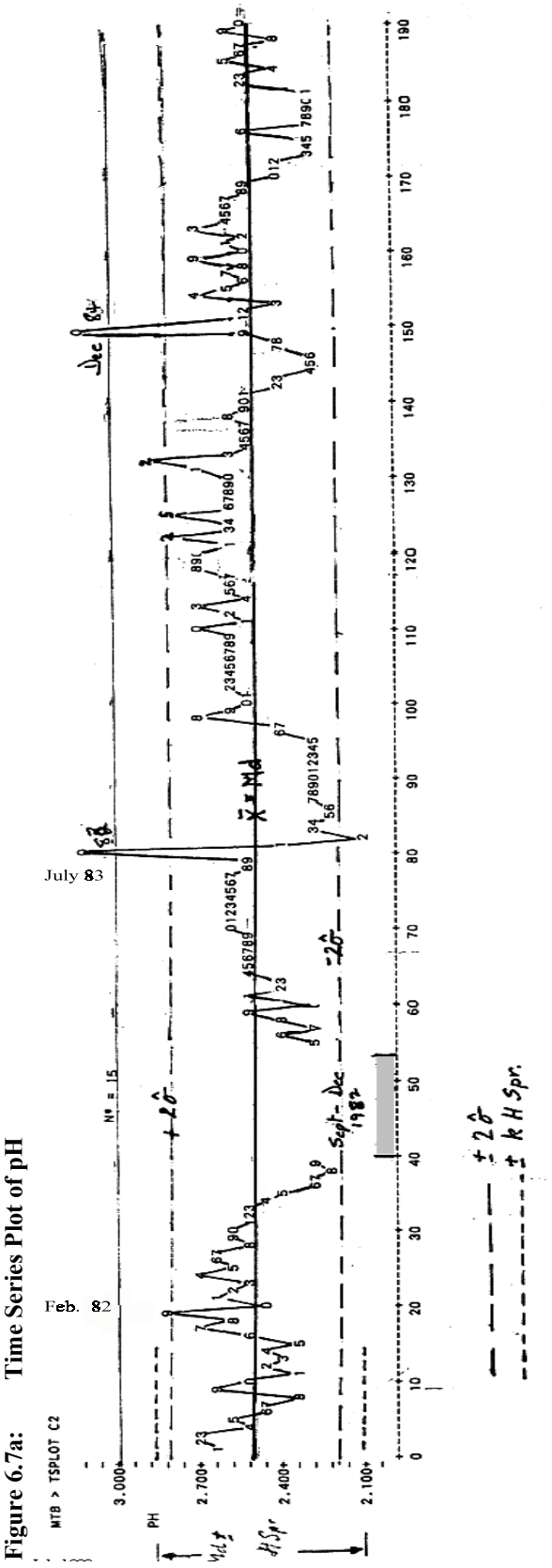


Figure 6.7a: Time Series Plot of pH

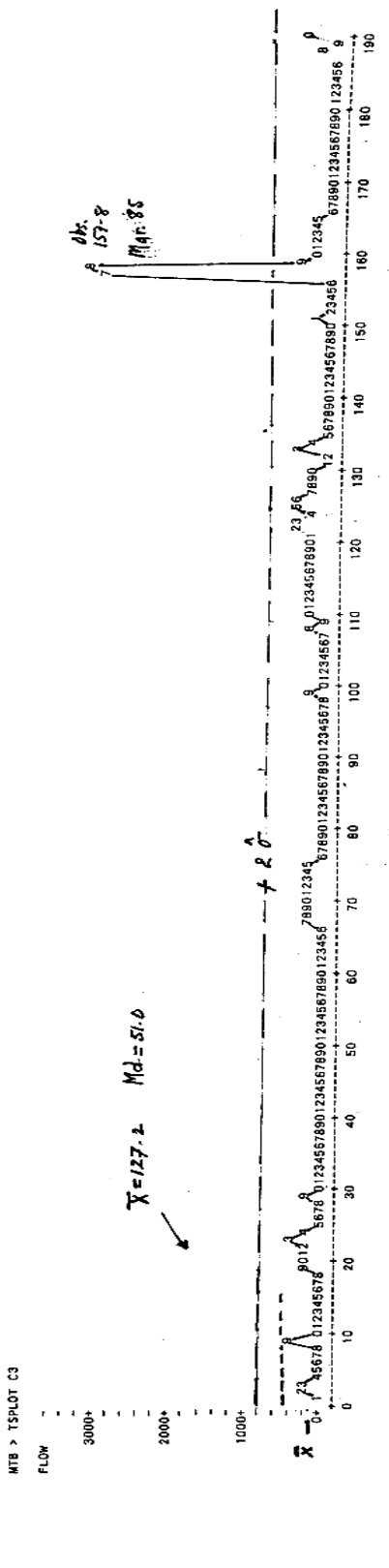


Figure 6.7b: Time Series Plot of Flow

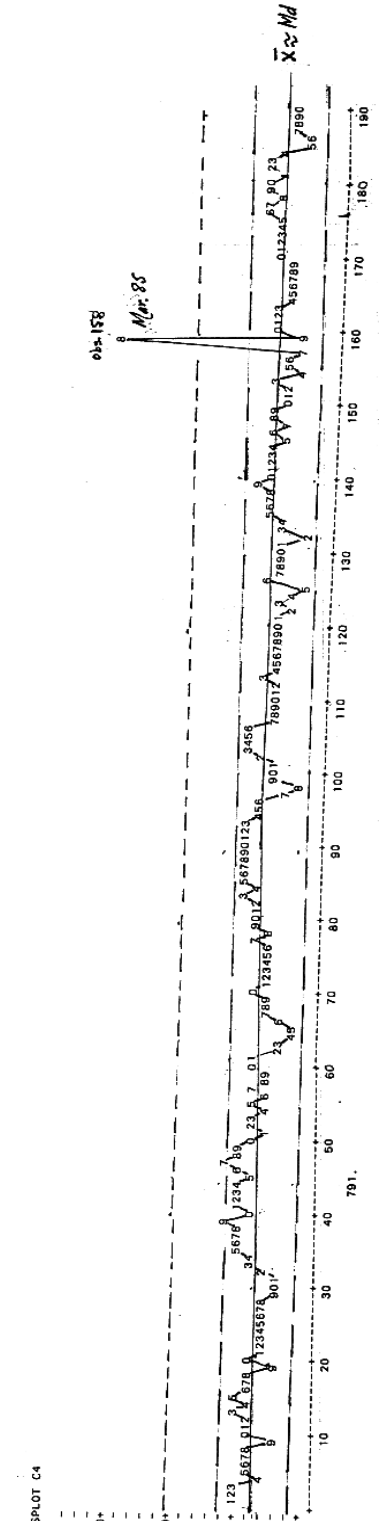


Figure 6.7c: Time Series Plot of Acidity

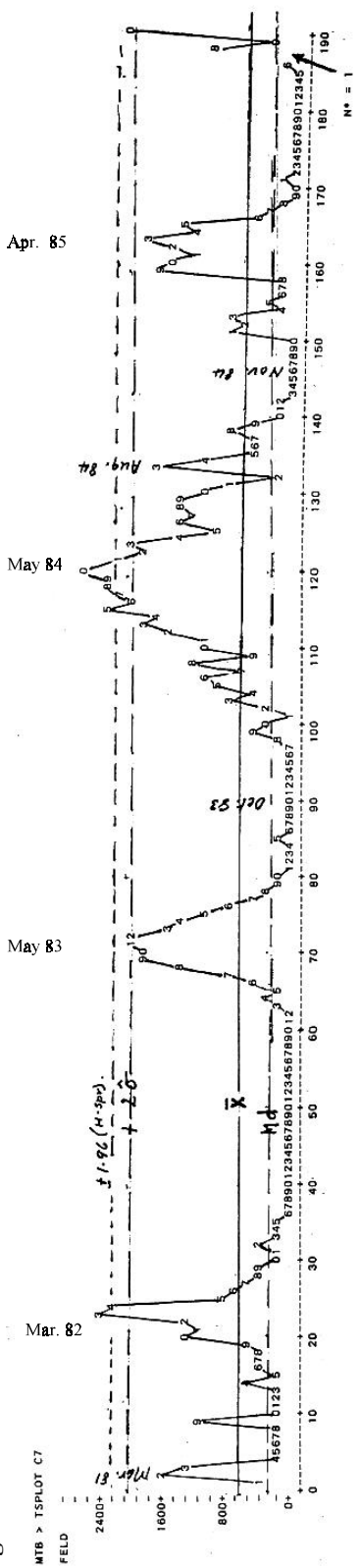


Figure 6.7d: Time Series Plot of Iron Load

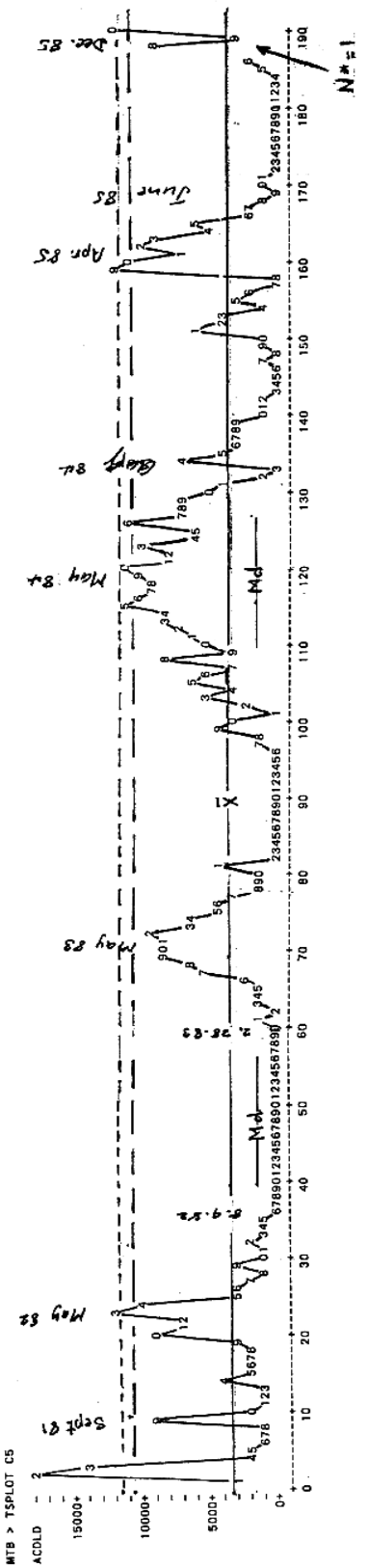


Figure 6.7e: Time Series Plot of Acid Load

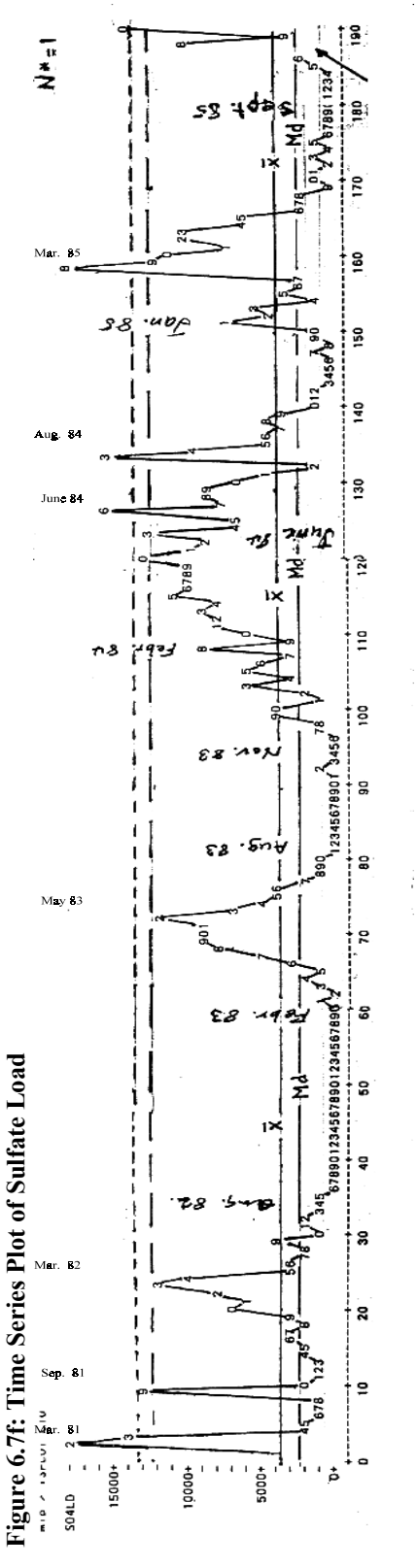


Figure 6.7f: Time Series Plot of Sulfate Load

Quality Control Limits for the Variables

Two measures of quality control were used to illustrate this aspect of the analysis. The first is conventional (mean ± 2x the standard deviation). The second is non-parametric (median ± 1.96 x a function of the H-spread). Since sample size, $N' = 1$, the function is: (1.25 * H-Spread / 1.35). Both measures are based on analysis of the Clarion data ("Quality Control Limits," Chapter 5). Summary statistics for these measures are listed in Tables 6.4 and 6.5. At the base of each table are statistics for the three load variables expressed in logarithms.

Table 6.4: Base Data for Calculation of Quality Control Limits of Ernest Data

No.	Variable	Mean $\overline{X}$	Median	R = H-spread	C-spread	$\hat{\sigma}$	H-spread/1.345
2.	pH	2.506	2.50	0.2	0.530	0.1524	0.148
3.	Flow	127.2	51.0	153.0	342.0	337.0	113.4
4.	Acid	3621.	3539.	1283.	3493.	1357.	951.1
5.	Acid Load	3367.	1843.	5210.	11307.	3639.	3862.1
6.	Total Iron	527.2	515.5	358.	647.	210.	265.4
7.	Iron Load	626.6	275.0	1096.	2193.	722.3	812.5
8.	Ferrous Iron	364.8	360.5	348.	807.	251.9	258.
9.	SO_4	3887.4	3804.	1583.	3933.	1105.2	1173.5
10.	SO_4 Load	3837.	2108.	5857.	13711.	4198.	4341.7
				Log. Data			
5.	Log Acid Load	3.120	3.265	1.127	1.869	0.631	0.835
7.	Log Iron Load	2.277	2.439	1.352	2.168	0.747	1.002
10.	Log SO_4 Load	3.175	3.324	1.089	1.917	0.632	0.807

Table 6.5: Two Measures of Quality Control (1) $\pm 2\ \hat{\sigma}$

(2) 1.96 [(1.25 H-spread) / $1.35\sqrt{N'}$]

No.	Variable	Mean ($\overline{X}$)	$\overline{X}$ ± 2 $\hat{\sigma}$	Md ± 1.96(..)	1.96 [(1.25 H-spread)/ 1.35 $\sqrt{N'}$	Median	2 $\hat{\sigma}$
2.	pH	2.506	2.201 to 2.811	2.137 to 2.863	0.363	2.500	0.305
3.	Flow	127.2	-546.8 to 801.2	-226.7 to 328.7	277.7	51.0	674.0
4.	Acid	3621	907 to 6335	1211 to 5867	2328	3539	2714
5.	Acid Load	3367	-3911 to 10645	-7612 to 11298	9455	1843	7278
6.	Total Iron	527.2	107.2 to 947.2	-134.2 to 1165.2	649.7	515.5	420.0
7.	Iron Load	626.6	-818.0 to 2071.2	-1714.0 to 2264.0	1989.0	275.0	1444.6
8.	Ferrous Iron	365	-139 to 868.6	-271 to 992	632	361	504
9.	SO_4	3887.4	1677.0 to 6097.8	931.1 to 6676.9	2872.9	3804.0	2210.4
10.	SO_4 Load	3837	-4559.0 to 12233.0	-8521.4 to 12737.4	10629.4	2108.0	8396.0

No.	Variable	Mean ($\overline{X}$)	$\overline{X}$ ± 2 $\hat{\sigma}$	Md ± 1.96(..)	1.96 [(1.25 H-spread)/ 1.35 $\sqrt{N'}$	Median	2 $\hat{\sigma}$
Log. Data							
5.	Log Acid	3.120	1.858 to 4.382	1.220 to 5.310	2.045	3.265	1.262
7.	Log Iron	2.277	0.783 to 3.771	-0.015 to 4.893	2.454	2.439	1.494
10.	Log SO_4	3.175	1.911 to 4.439	1.348 to 5.300	1.976	3.324	1.264

These two quality control limits are inserted as dashed lines in Figures 6.7a through 6.7f. For pH (Figure 6.7a), similar limits resulted from both measures. The mean and median coincide (on the scale of the graph in Figure 6.7a), and frequency distribution of pH is essentially symmetrical.

For flow and acidity (Figures 6.7b and 6.7c respectively), the standard deviations are inflated by rare extreme values. Thus, the quality control limits for both measures are essentially insensitive except to the extremes. It should be noted that the range is small.

The quality control limits for the load variables are wide, with the lower limits falling below zero. The lower limits are, therefore, omitted from the graphs (Figures 6.7d through 6.7f). The respective means and medians are not very different in magnitude and neither are the positive control limits. The use of either quality control limit would have little effect in identifying exceedences of baseline pollution load. It appears from these graphs that either measure would suffice. Sensitivity to departures from set limits could be increased by dividing by the square root of N', or by increasing sample size (e.g., from 1 to 4). This would reduce the range to half its original value. On the other hand, use of the root N' factor with N' > 1, could increase the sensitivity too much and many values of these widely fluctuating parameters would fall outside the limits thereby calling for action. If fluctuations arise from "natural causes" and not from mining activity, this would be undesirable. Obviously, the entire range of pH, for example, is small (2.1 – 3.1) and the discharge is consistently acidic.

Model Identification

Autocorrelation functions form the basis for model identification in applying full-scale Box-Jenkins time series analysis. Hence, the autocorrelation and partial autocorrelation functions were run on the data for each variable. The graphs are presented in Figures 6.8a, 6.8b, 6.8c, 6.8e, and 6.8g, for the autocorrelation functions (Acf) and Figures 6.8d, 6.8f, and 6.8h for the partial autocorrelation functions (Pacf).

Figure 6.8a: Autocorrelation Function of pH

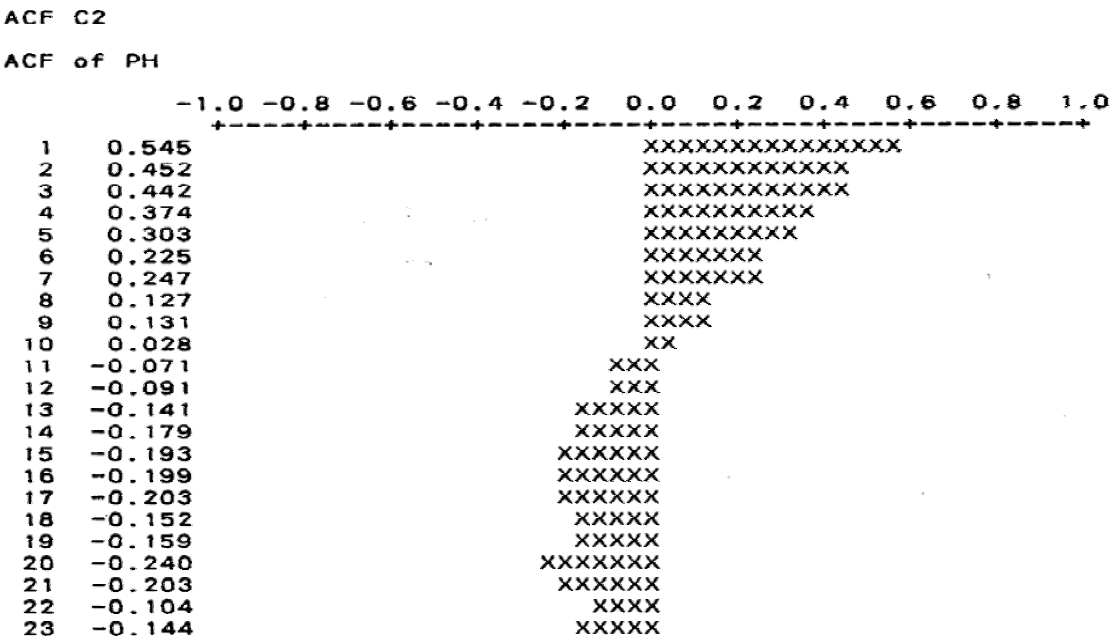


Figure 6.8b: Autocorrelation Function of Iron

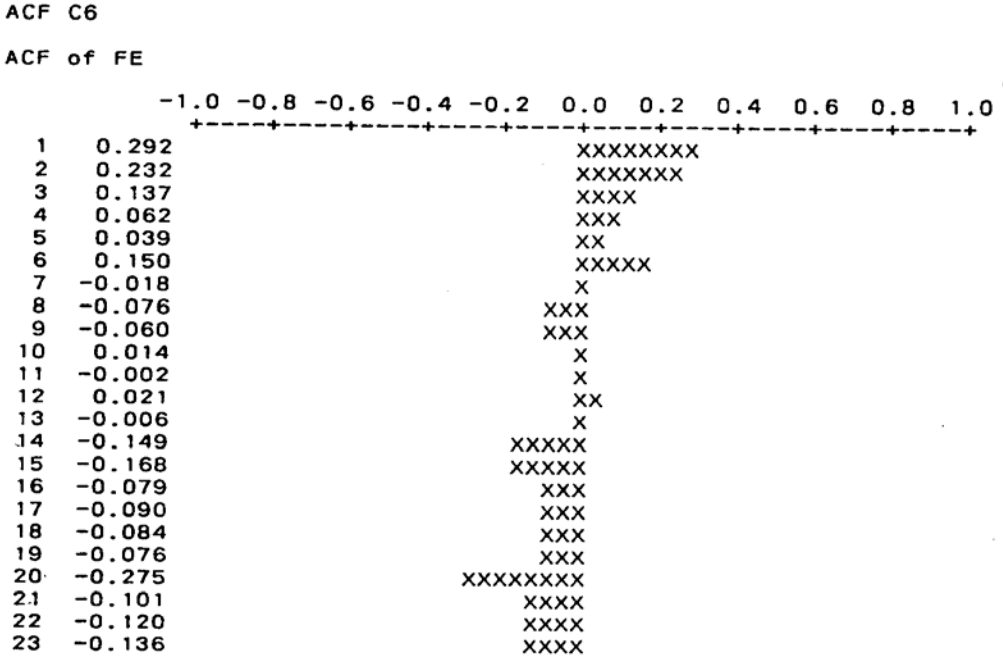


Figure 6.8c: Autocorrelation Function of Flow

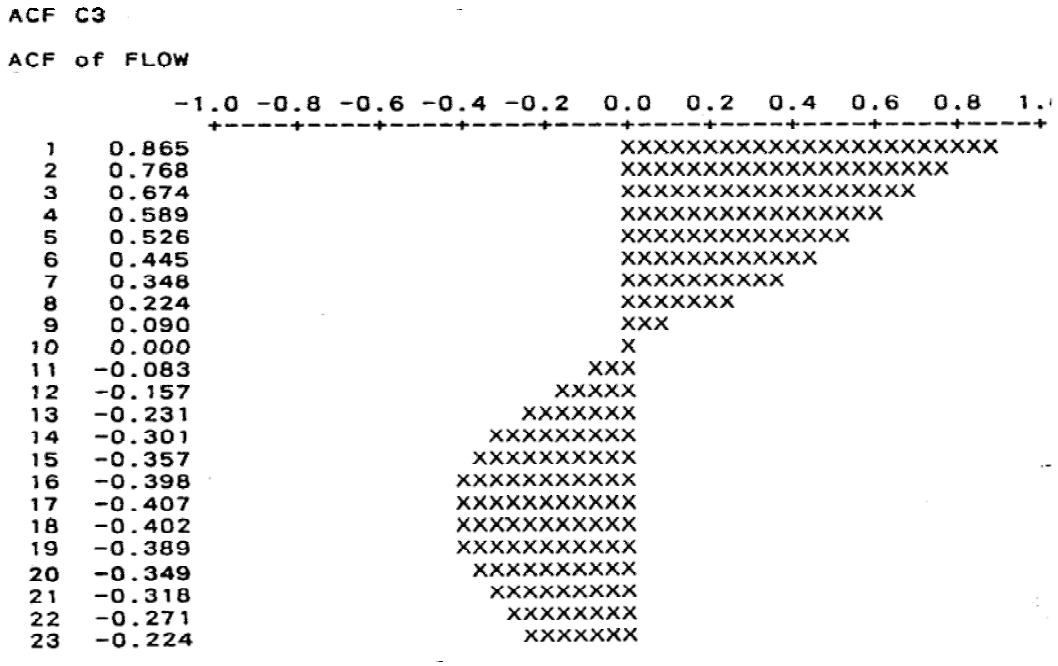


Figure 6.8d: Partial Autocorrelation Function of Flow

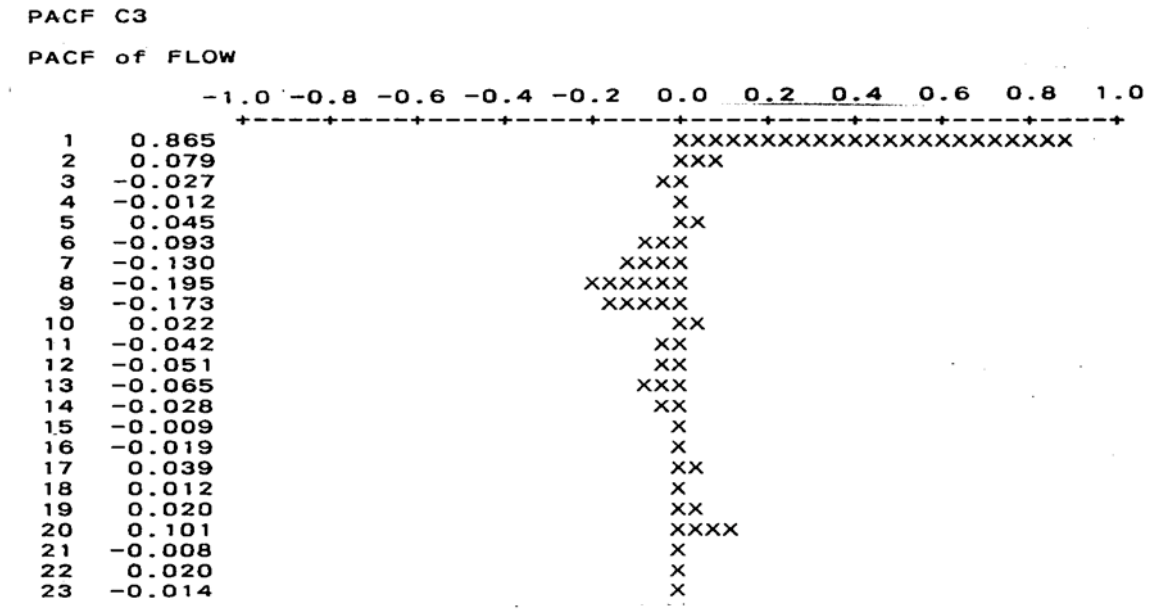


Figure 6.8e: Autocorrelation Function of Acidity

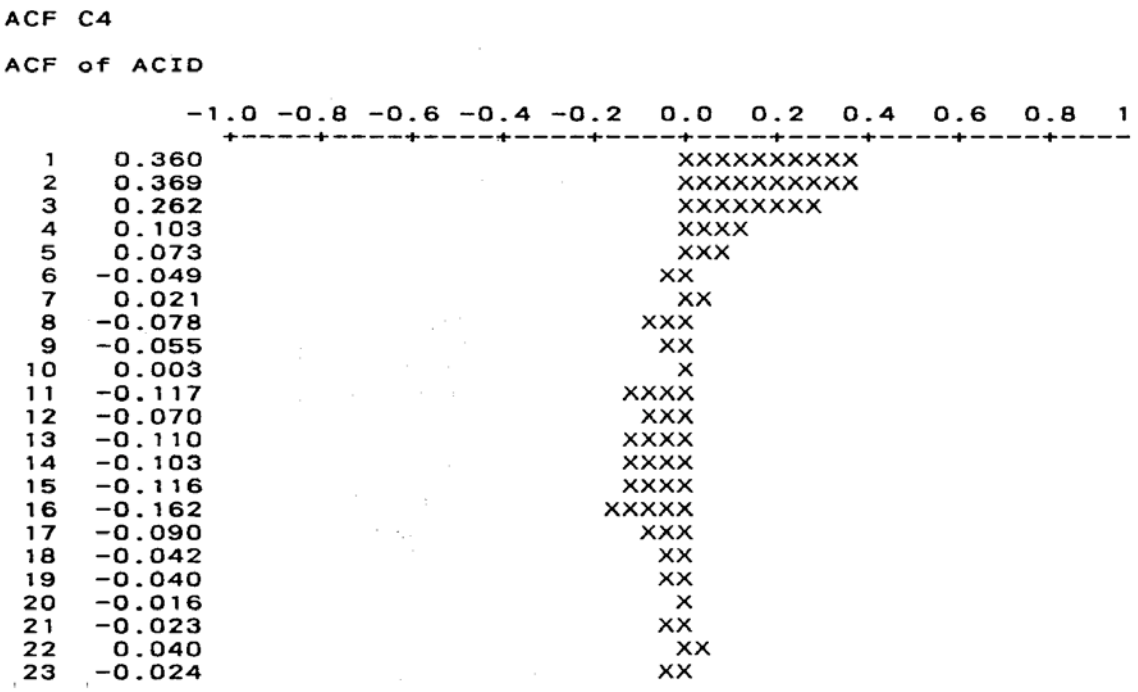


Figure 6.8f: Partial Autocorrelation Function of Acidity

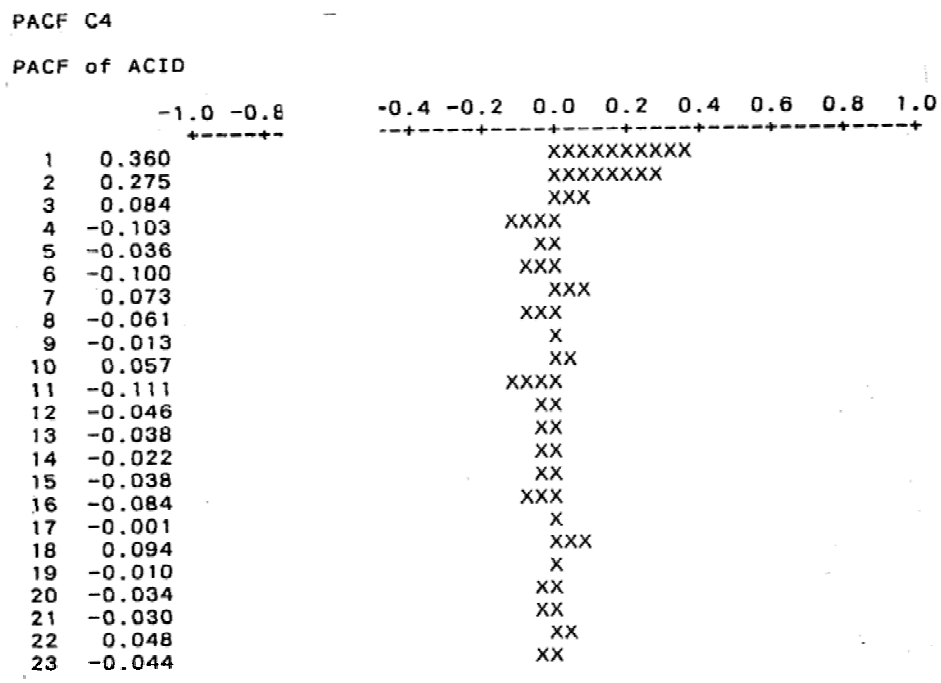


Figure 6.8g: Autocorrelation Function of Acid Load

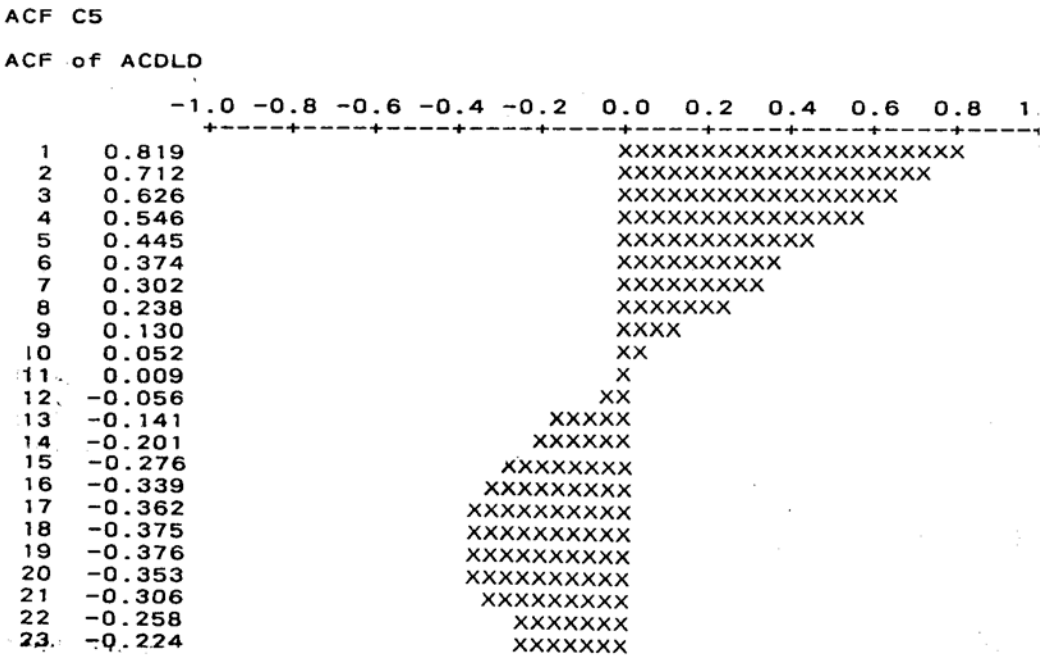
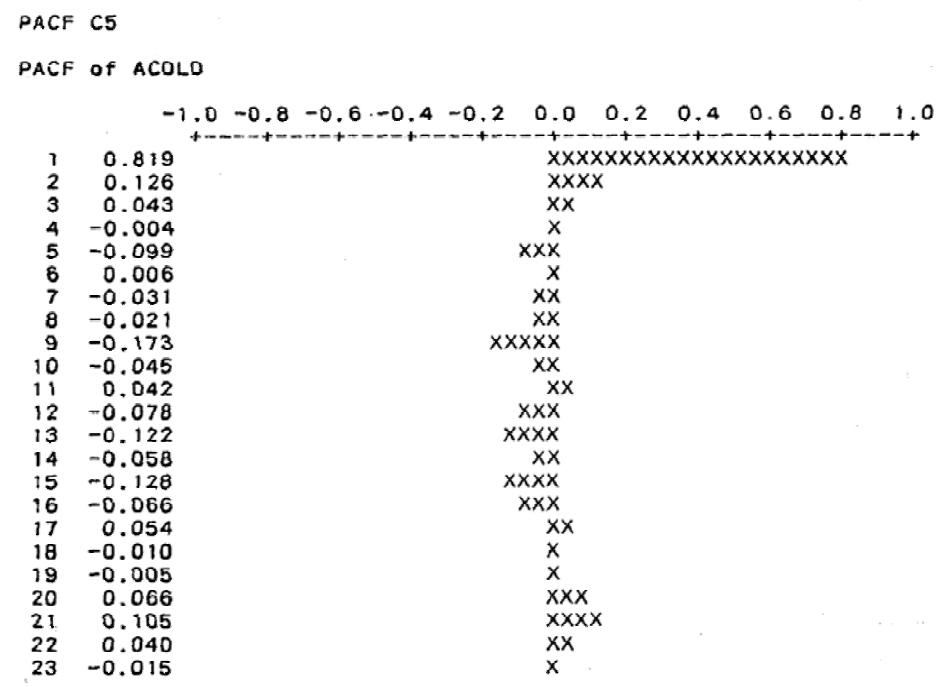


Figure 6.8h: Partial Autocorrelation Function of Acid Load



pH (Figure 6.8a), flow (Figure 6.8c), and three load parameters (e.g., see Figure 6.8g for acid load) yield similar Autocorrelation functions (Acf's). The concentration variables acidity (Figure 6.8e), iron (Figure 6.8b), and sulfate (Figure not available) also show similar Acf's, but the former set (which includes load) differs from the latter. The former set shows a strong decline throughout the function. This decline is confirmed by the single large spike at lag 1 in the corresponding partial autocorrelation factors (Pacf's, Figures 6.8d and 6.8h). This behavior implies that all these variables require a first difference to remove the trend. The Acf and Pacf for each of the concentration variables (e.g., Figures 6.8e and 6.8f) suggest moving average (MA) models with at most two terms (or one term and a first difference). It is perhaps advisable to try an auto-regressive moving average (ARMA) model in which the AR term could proxy for the first difference and the MA term would take care of the remainder.

Model Fitting: pH

It was decided to attempt to fit an auto-regressive integrated moving average (ARIMA) model (1,1,1) to variation in pH. The correlation coefficient between the AR and MA coefficients was r = 0.81, which implies that they are closely associated (i.e., both are unlikely to be necessary). Testing the Acf of the residuals yielded a chi-square = 27.16 with 28 degrees of freedom (i.e., the Acf is not significantly different from that of white noise). There is only one significant spike at lag 20 in this Acf, thus, it is effectively clean. Any further differencing results in overdifferencing (i.e., chi-square of the Acf increases to significant again). The model has improved the variation (Pacf of the residuals has no significant spikes) but contains an unnecessary coefficient $\hat{\Phi}_1$. Clearly, the AR (1) is adequately taken care of by the first difference.

If we now fit a moving average model with a first difference (i.e., an MA (0,1,1) model), the Acf of the residuals yields a chi-square of 26.87 with 29 degrees of freedom (thus, not significantly different from an Acf of white noise). Any further differencing overcompensates. The only significant spike is at lag 20 as in the previous model. Because this is an isolated significant autocorrelation way out from zero lag, it is considered a random discrepancy. The Pacf of the residuals is also clean. The 95% confidence limits around the MA coefficient ($\hat{\theta}_1$) does not contain zero. Hence, the MA coefficient is significantly different from zero (real) and, incidentally, about the same size as in the ARIMA model ($\hat{\theta}_1$ = 0.594). The residual standard deviation is $\hat{\sigma}_e$ = 0.126, a reduction in the pH of the original data from 0.152 to 0.126. The relationship may be expressed as:

$$z_t = z_{t-1} + a_t - 0.594a_{t-1}$$

Model Fitting: Flow (Log)

An AR (1,0,0) model was fitted to the variation in logarithms of the flow variable; it was considered that the AR(1) coefficient would "take care of" the first difference. Chi-square of the residual = 30.06 with 28 degrees of freedom (0.50 > P > 0.30; i.e., not significantly different from that expected from white noise). There are no significant spikes in the Acf or Pacf values.

This model yields the following equation with standard deviation of the residuals $\hat{\sigma}_e = 0.347$ (reduced from 0.697 for the original standard deviation of the logarithms of flow in Table 6.2):

$$z_t = 0.873\, z_{t-1} + 1.636 + a_t$$

Model Fitting: Acidity (Log)

From the Acf, developed during the identification step of the Box-Jenkins series, it was decided to try an MA (0,1,2) model which would presumably clear out the large spikes at the first three lags in the Acf. Upon fitting, it turned out that the correlation coefficient between the two moving average coefficients ($\hat{\theta}_1$ and $\hat{\theta}_2$) was −0.612 (i.e., as one increased the other decreased). A chi-squared test of the residual Acf yielded 29.86 with 28 degrees of freedom (0.50 > P > 0.30). The Acf spike at lag 6 is significantly larger than its error.

Upon testing the coefficients of this model, the $\hat{\theta}_1 = 0.642$ and is real, but the second $\hat{\theta}_2 =$ −0.640 and its confidence belt included zero. The standard deviation of the residuals is 0.139.

An MA (0,0,2) model showed no correlation among the two coefficients or between either coefficient and the mean. The residual chi-square = 32.77, with 27 degrees of freedom (0.30 > P > 0.20) is not significantly different from that expected from white noise (random error). The relevant equation is:

$$z_t = 3.536 + a_t + 0.205a_{t-1} + 0.274a_{t-2}$$

with standard deviation of the residuals as $\hat{\sigma}_e = 0.136$, a small improvement over the MA (0,1,2) model and some slight improvement over the original standard deviation (0.147) of the variable logarithms given in Table 6.2.

Model Fitting: Acid Load (Log)

As a first approximation, an MA (0,1,1) was fitted to these data and a trend term was included to determine if it gave rise to any improvement. The Acf of the residuals yielded a chi-square = 22.41 with 28 degrees of freedom (0.80 > P > 0.70), not significantly different from an Acf of white noise. A barely significant spike occurred at lag 16 in the Acf and Pacf. It was not supported by any other diagnostic characteristic and so was ignored. The correlation coefficient

between the trend constant and the MA coefficient ($\hat{\theta}_1$)= −0.01. Therefore, they are effectively independent. However, on testing the trend term, its 95% confidence limits include zero, and therefore, the trend constant does not make any real contribution to explaining the variation of log acid load. The MA coefficient ($\hat{\theta}_1$) = 0.247 and is real. The equation may be expressed as (the trend term is omitted for reasons given above):

$$z_t = a_t - 0.247a_{t-1}$$

The standard deviation of the residuals is 0.355, which is approximately half the original standard deviation of 0.614.

Two other models were fitted to these data (an ARI (1,1,0) and an ARMA (1,0,1)), again assuming that the AR coefficient would proxy for the first difference in the ARMA model. A chi-square of the residuals from the ARI model yielded 22.11 with 29 degrees of freedom (0.90 > P > 0.80). Clearly, the first differences and the autoregressive coefficient (Φ_1) reduced any unusual occurrences in the data. There were no significant spikes in the Acf but there is a possible one at lag 16 in the Pacf (i.e. the MA (0,1,1) model). The AR coefficient was significantly different from zero ($\hat{\Phi}$ = −0.203) and the standard deviation of the residuals is $\hat{\sigma}_e$ = 0.353, a considerable reduction from the original value of 0.614 for standard deviation of the logarithms (see Table 6.2). The equation is:

$$z_t = 0.797z_{t-1} - 0.203z_{t-2} + a_t$$

The ARMA (1,0,1) model possessed two coefficients and a mean. Their respective correlations were r_{12} ($\hat{\Phi}_1$ vs. $\overline{X}$) = 0.03, r_{13} = 0.55, and r_{23} ($\overline{X}$ vs. $\hat{\theta}_1$) = 0.01, effectively independent for the first and third and not very large for the second. Acf of the residuals yielded a chi-square of 24.32 with 27 degrees of freedom (0.70 > P > 0.50), indicating no significant difference from an Acf for white noise. The autoregressive coefficient ($\hat{\Phi}_1$ = 0.881) and the mean ($\overline{X}$ = 3.196) were real, whereas the 95% confidence limits around the moving average coefficient ($\hat{\theta}_1$ = 0.171) contains zero. The standard deviation of the residuals is 0.347, the same order of magnitude as the previous models fitted to log acid load.

Summary

It is somewhat surprising that there appears to be no seasonal component in the time series models, particularly in the load variables. The only satisfactory explanation appears to be the existence of too many maxima at too many different times with very little repetition during the same time period.

Most of the variables show the presence of a trend over time (pH, flow, acidity, acid load, iron load, ferrous iron). These variables need a first difference to remove the effects of the trend. It

seems evident from the studies to date that a moving average model applied to the first differences is almost universally the best choice. In some cases, the autoregressive model, possibly with a first difference, is also appropriate. In both cases, there is an indicator that the variation in whichever parameter is being analyzed, when first differenced, leads to a random walk.

The quality control analysis, in both cases, suggests that either the mean (plus or minus two standard deviations) or the non-parametric median (plus or minus a function of the H-spread) are equally appropriate. For the present, it is recommended both should be used until one or the other show superior performance.